Artificial Intelligence in Data Processing

Alexander Measure
Economist
Office of Safety and Health Statistics
FedCASIC 2013

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

www.bls.gov

Survey of Occupational Injuries and Illnesses (SOII)

- Annual Survey
- 250,000 cases
- Text narratives for:
  - Occupation
  - Injury/Illness Characteristics

BLS

Example Case

Job title: janitor

What was the employee doing just before the incident?
mopping floor in gym

What happened?
slipped on wet floor and fell

What part of the body was affected?
fractured right arm

What object directly harmed the employee?
wet floor

Occup: 37-2011 (Janitor)

Nature: 111 (Fracture)

Event: 422 (Fall, slipping)

Part: 420 (Arm)

Source: 6620 (Floor)

BLS

Limitations of Human Coding

- Time consuming
- Very difficult
  - Detailed classifications
  - Ambiguous data
- Inconsistent coding

BLS

Artificial Intelligence

- Text Classification
  - Rules
  - Machine Learning
- Ex. Census
  - Rule based system: 192 person-months
  - Machine learning: 4 person-months
    - Source: Creecy et al., 1992

BLS

Machine Learning

- 3 Steps
  1. Select feature representation
  2. Select model
  3. Fit model to data

BLS

Feature Representation

- Each feature corresponds to a word
- Job title: “assistant nurse”
  - $X_{nurse} = 1$
  - $X_{assistant} = 1$
  - $X_{mechanic} = 0$

BLS

Logistic Regression Model

$$P(\text{code} = \text{A}) = \frac{\exp(w_{a1}x_{\text{nurse}} + w_{a2}x_{\text{assistant}} + w_{a3}x_{\text{mechanic}})}{Z}$$

$$P(\text{code} = \text{B}) = \frac{\exp(w_{b1}x_{\text{nurse}} + w_{b2}x_{\text{assistant}} + w_{b3}x_{\text{mechanic}})}{Z}$$

$$P(\text{code} = \text{C}) = \frac{\exp(w_{c1}x_{\text{nurse}} + w_{c2}x_{\text{assistant}} + w_{c3}x_{\text{mechanic}})}{Z}$$

BLS

Autocode

- New Case: “assistant mechanic”
  - $X_{nurse} = 0$
  - $X_{assistant} = 1$
  - $X_{mechanic} = 1$

- P(code=A) = .05
- P(code=B) = .25
- P(code=C) = .70 ←

BLS

Does it work?

- Train on 250k
- Test on 10k

	Computer to Original (%)
Occupation	80
Part	81
Nature	80
Event	49
Source	60

Humans vs. Computer

- 1000 cases
- 3 re-coders

	Human to Original (%)	Computer to Original (%)
Occupation	66	80
Part	82	81
Nature	76	80
Event	47	49
Source	56	60

BLS

Limitations

- Training data
- New words
- Linguistic complexity

Applications

- Review
- Assisted Coding
- Partial Autocoding

BLS

Useful Resources

- Free Machine Learning Class
  - https://www.**coursera**.org/course/ml
- Free Software
  - Python (Scikit-Learn)
- Free Support
  - MetaOptimize Q&A
  - Stack Overflow

BLS

Contact Information

Alexander Measure

Economist

Office of Safety and Health Statistics

www.bls.gov/iif

202-691-6185

measure.alex@bls.gov

BLS

BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR